Additional File 1

Some study design considerations

The amplification protocol.

In an ideal UMI-based protocol, the UMI-bearing primers would be used for just one round of amplification and then removed to ensure that templates were not sampled more than once, and that UMIs were not switched in later rounds of amplification. While such a protocol is possible (e.g. Salipante et al. 2013), the manipulations involved would be arduous for a study involving a large number of samples, and would risk losing samples at a stage when the DNA concentration is still very low. Instead, we opted for a straightforward protocol using a "one-pot" initiation and amplification system. Forward primers consisted of two modules; an inner primer bearing the UMI and designed to amplify the target gene, and a universal outer primer that binds only to a linker on the inner primer. To minimise the participation of the inner primer in later rounds of amplification, it is added at one hundredth of the concentration of the outer primer.

Choice of amplicons and primers.

We chose genes that were known to be suitably polymorphic in the target species, and used available sequence data to design primers which amplify all known variants of the genes within the target species, but exclude the orthologous genes in related species as well as any paralogs in the genome. All primers were adjusted to have similar predicted melting temperatures and to avoid problematic secondary structures, using standard primer design tools. Primer pairs were chosen to produce amplicons of similar length for each gene, and this length was short enough to allow the paired-end sequencing to cover both strands completely in order to maximise sequencing accuracy. We observed that the targets varied in efficiency of amplification, emphasising that the primer pairs should be tested individually prior to undertaking the study. In principle, it should be possible to multiplex the amplification of several target genes in one reaction, but we chose to amplify the genes separately so that we could check that all genes were well represented. The genes of choice should be verified to be single copy in the genome. Unknown to us, some strains of our target species had a second copy of *nodD*, including the strain SM170C that we used for the synthetic mixture. Despite having a lower primer specificity than the target *nodD* sequence, the second copy was amplified and this hindered the downstream analysis since there were three 'genuine' sequences rather than the expected two.

Length of the UMI sequence.

The probability of assigning the same UMI to more than one sequence by chance is higher than one might expect - the "birthday paradox" (Sheward et al., 2012). Earlier studies used UMIs of 8 or 10bp (Brodin et al., 2015; Jabara et al., 2011), but as the read length of high-throughput sequencing has increased, there is less constraint to keep the UMI so short. In this study, we have used a 12bp UMI to increase the number of possible sequences to ~4.5 million, which will greatly reduce, though not eliminate, collisions. Guanine is excluded at certain positions to reduce potential secondary-structure problems.

Choice of polymerase: sensitivity versus accuracy.

We used both Platinum Taq and Phusion High Fidelity polymerase, which is reported to have a lower error rate (Kinde et al., 2011). Nichols et al. (2018) tested a wider range of

polymerases for metabarcoding and also found variation in error rate, as well as in sensitivity to template GC content. Our results confirm that Phusion generates fewer errors, including fewer chimeras, so we recommend that a high-fidelity enzyme is used when template concentration is high, as in our root nodule extracts. However, we originally developed the MAUI-seq approach to study the diversity of *Rhizobium nodD* sequences in soil, where the template DNA is in very low abundance, and we found that amplification could be obtained reliably with Platinum but not with Phusion (Boivin et al., 2020). Therefore, for samples limited by low DNA yield, it may be necessary to use a more sensitive but more error-prone polymerase. In these circumstances, the error correction provided by the UMIs becomes even more advantageous.

MAUI-seq laboratory step-by-step protocol

1.1 Field sampling and DNA extraction of nodule samples

- Collect and wash White clover (*Trifolium repens*) roots.
- Collect 100 large pink nodules from 4 points on each plot. Store nodules at -20°C until DNA extraction.
- Thaw nodule samples at ambient temperature and crush using a sterile homogeniser stick. Mix crushed nodules with 750µl Bead Solution from the DNeasy PowerLyzer PowerSoil DNA isolation kit (Cat No./ID: 12855-100, QIAGEN, USA).
- Extract DNA from the nodule samples following the manufacturer's instructions.
- DNA sample concentrations can be calculated using Nanodrop 3300 (Thermofisher Scientific Inc., USA).

1.2 DNA extraction of in vitro cultured samples

- Grow strains on Tryptone Yeast agar (28°C, 48hrs).
- Resuspend culture in 750µl of the DNeasy Powerlyzer PowerSoil DNA isolation kit (Catalog No.: 12855-100, QIAGEN, USA).
- Extract DNA following the manufacturer's instructions.
- Calculate DNA sample concentrations using QuBit (Thermofisher Scientific Inc., USA). Dilute DNA samples of the two strains to the same concentration and mix in various ratios (**Supplementary Table 1**).

2. MAUI-seq PCR and AMPure XP Bead Purification

- Primer sequences were designed for two *Rlt* housekeeping genes, recombinase A (*recA*) and RNA polymerase B (*rpoB*) and for two *Trifolium repens* specific *Rhizobium* symbiosis genes, *nodA* and *nodD* (**Table S1**). Primers include: a forward gene specific inner primer, a universal forward outer primer, and a reverse gene specific primer.

Table S1. Primer sequences for MAUI-seq PCR.

Primer-type	Primer name	Sequence
Forward gene specific inner primer	QQMf1-rpoB-Inner	AGATGTGTATAAGAGACAG-A–NNNHNNNWNNNH-GyTCGCAGTGGTGGATGTT
	QQMf1-recA-Inner	AGATGTGTATAAGAGACAG-A–NNNHNNNWNNNH-CGAGAATGTTGTCGAGATyGAGACGA
	QQMf1-nodAt-Inner	AGATGTGTATAAGAGACAG-A–NNNHNNNWNNNH-CCGGATCTsGAGGGGCT
	QQMf1-nodDt-Inner	AGATGTGTATAAGAGACAG-A–NNNHNNNWNNNH-ATGCGTTTTAAGGGmyTGGATCT
Forward universal outer primer	QQMf1-Outer	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-A
Reverse gene specific primer	QQMr1-rpoB	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG TCCGTCTTCRAGGAACGGCAT
	QQMr1-recA	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG TTATCGGTGATTTCRAGCGCCTG
	QQMr1-nodAt	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAC TGCANCCGTTTCGTTCGATCAATGA
	QQMr1-nodDt	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG TGCRCCGGTCAGATTCCGC

- PCRs are undertaken individually for each primer set using Platinum Taq DNA polymerase (Catalog No.: 10966018, Thermofisher Scientific Inc., USA) using the following PCR reaction mixture and thermocycler programme (**Table S2, Table S3**).
- The PCRs can also be undertaken using the Phusion High-Fidelity DNA polymerase (Catalog No.: F530S, Thermofisher Scientific Inc., USA) (**Table S4**) using the same thermocycler progamme (**Table S3**).

Table S2. MAUI-seq PCR reaction mixture for Platinum Taq DNA polymerase.

	PCR recipe for 25µl reaction	Initial concentra-tion	Final concentra-tion	
Water	15.9			
buffer	2.5	10	1	
MgCl2	1.5	50	3	mM
dNTP	0.5	10	0.2	mM
QQMf1-*recA** (specific gene inner primer)	1	0.1	0.004**	µM
QQMf1 (universal outer primer)	1	10	0.4	µM
QQMr1-*recA** (this is the full length reverse primer specific to recA – no inner primer)	1	10	0.4	µM
Platinum HS TAQ	0.1	5	0.02	Units/µl
Template DNA	1.5			

**recA* is used as an example gene. **Concentration used in this study. The optimal concentration will be dependent on the amplicon. In our case a lower concentration would, in retrospect, have been preferable.

Table S3. MAUI-seq PCR programme (approximately 2 hours 20 mins total).

Temperature (°C)	Time (seconds)	Cycles
95	180	
94	30	
70	300	2
72	120	
94	30	
70	60	30
72	60	
72	600	
4	hold	

Table S4. MAUI-seq PCR reaction mixture for Phusion High-Fidelity Taq proof-reading DNA polymerase. **recA* is used as an example gene.

	PCR recipe for 25ul reaction	Initial concentration	Final concentration	
Water	14.75			
5X Fusion HF Buffer	5			
dNTP	0.5			
QQMf1-*recA** (specific gene inner primer)	1	0.1	0.004	µM
QQMf1 (universal outer primer)	1	10	0.4	µM
QQMr1-*recA** (this is the full length reverse primer specific to *recA* – no inner primer)	1	10	0.4	µM
Phusion Polymerase	0.25			
Template DNA	1.5			

- Perform PCRs individually for the four primer sets to ensure an equal amount of product is produced for each gene. A no-template-control and a positive control (DNA extract from another known *Rlt* nodule sample) is included in each PCR run. Therefore, 4 PCR products are produced for each sample.
- Successful PCR amplification can be confirmed on a 0.5X TBE 2% agarose gel run at 90V for 2 hours (expected band size: 381bp).
- Pool the four PCRs for each sample (four 20µl PCR reactions for each sample yield a total of 80µl for each sample).
- Purify samples using AMPure XP Beads following the manufacturer instructions (**Table S5**) (Beckman Coulter, USA).

Table S5. AMPure XP Bead Purification protocol

Step	AMPure XP bead purification protocol
1	Vortex beads at room temperature until they are completely resuspended
2	For 20µl PCR product add 16µl beads (For 80µl PCR product use 64µl of beads)
3	Pipette up and down 10 times
4	Incubate at ambient/room temperature for 5 minutes, allowing DNA to bind to beads
5	The following steps should be carried out on a magnetic stand: (a) Place samples on a magnetic stand for 2 minutes or until supernatant has cleared (b) Remove supernatant (c) Add 200µl freshly made 80% ethanol (d) Incubate for 30 seconds at room temperature (e) Remove supernatant (f) Repeat step c to e (g) Air dry for 10 minutes
6.	Remove tubes from magnetic stand
7.	Add 52µl 10 mM Tris pH 8.5/nuclease free water to elute DNA from beads
8.	Pipette up and down 10 times to fully resuspend beads
9.	Incubate for 2 minutes at room temperature
10.	Return to magnetic stand for 2 minutes
11.	Carefully transfer 50µl supernatant to a new sterile tube without disturbing the pellet

3. Nextera XT indexing for multiplexing and sequencing (Table S6 and Table S7)

- Index samples for multiplexing sequencing libraries with Nextera XT DNA Library Preparation Kit v2 set A (Catalog No.: FC-131-2001, Illumina, USA) using the Phusion High-Fidelity DNA polymerase (Catalog No.: F530S, Thermofisher Scientific Inc., USA).
- The components for 50µl reaction volume and PCR programme are specified in Table 6 and 7, respectively. Indices should be added in unique combinations as specified in the manufacturer instructions (Illumina, USA).
- For the Nextera XT indexing PCR, make up the master mix first containing water, buffer, dNTP and Taq, and leave on ice.
- Then to each tube add the respective 5µl XT index S, and then 5µl XT index N so that each sample has a unique indices combination.
- Add 5µl of DNA PCR sample.
- Finally, add 35µl of the mastermix to each tube.

Table S6. Nextera XT indexing PCR reaction mixture for Phusion High-Fidelity Taq proof-reading DNA polymerase.

	PCR recipe for 50µl reaction
Water	23.5
Buffer	10
dNTP	1
XT index S	5
XT index N	5
Phusion	0.5
DNA	5

Table S7. Nextera XT index PCR programme.

Temperature (°C)	Time (seconds)	Cycles
95	180	
95	30	
55	30	10
72	30	
72	300	
4	hold	

4. Gel purification and sequencing

- Depending on the extent of spurious DNA bands present from indexing, gel extraction may not be required. If gel extraction is not required, PCR product can be cleaned with AMPure XP beads instead, as previously

- The PCR product can be purified on a 0.5X TBE 1.5% agarose gel and extracted with the QIAQuick gel extraction kit (Catalog No.: 28704, QIAGEN, USA). Expected band length after addition of Nextera XT indices: ~454bp.
- Load 10µl PCR products from the Nextera XT PCR onto a 1.5% agarose gel to purify the product prior to sequencing.
- Run the gel for 2 hours at 80V. Gel wells should be loaded with an empty well between each sample, to prevent product transfer between wells.
- Gel extraction should be carried out following the manufacturer's instructions, with the exception of:
    - Step 9: Add 30µl Buffer EB heated to 50°C and incubate for 5 minutes prior to centrifugation.
- Gel analysis: Use 2µl loading dye and 2µl gel purified DNA on normal 2% agarose gel. This was used to normalise the DNA concentration of the samples prior to sequencing.
- Normalise PCR amplicon concentrations to 10nM before MiSeq sequencing by visualising gel band intensities using GelAnalyzer2010a (Lazar, 2010).
- A pooled sample was quantified, quality checked and 2x300nt paired-end Illumina MiSeq sequenced by the University of York Technology Facility.

References

Boivin S, Lahmidi NA, Sherlock D, Bonhomme M, Dijon D, Heulin-Gotty K, Le-Queré A, Pervent M, Tauzin M, Carlsson G, Jensen E, Journet E-P, Lopez-Bellido R, Seidenglanz M, Marinkovic J, Colella S, Brunel B, Young P, Lepetit M. (2020) Host-specific competitiveness to form nodules in *Rhizobium leguminosarum* symbiovar *viciae*. New Phytologist. DOI: 10.1111/nph.16392

Brodin, J., Hedskog, C., Heddini, A., Benard, E., Neher, R. A., Mild, M., & Albert, J. (2015). Challenges with using primer IDs to improve accuracy of next generation sequencing. PLoS One, 10(3), e0119123.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). Cd-hit: accelerated for clustering the next generation sequencing data. Bioinformatics, 28(23):3150–3152.

Jabara, C.B. et al., 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proceedings of the National Academy of Sciences of the United States of America, 108(50), pp.20166–20171.

Kinde, I. et al., 2011. Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of the National Academy of Sciences of the United States of America, 108(23), pp.9530–9535.

Lazar, I. (2010). Gelanalyzer 2010a: Freeware 1d gel electrophoresis image analysis software. Available at: http://www.gelanalyzer.com/.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22:1658–9.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cow- ley, A. P., and Lopez, R. (2013). Analysis tool web services from the embl-ebi. Nucleic acids research, 41(W1):W597–W600.

Nichols, R.V., Vollmers, C., Newsom, L.A., Wang, Y., Heintzman, P.D., Leighton, M., Green, R.E. and Shapiro, B., 2018. Minimizing polymerase biases in metabarcoding. Molecular ecology resources, 18(5), pp.927-939.

R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Salipante, S. J., Sengupta, D. J., Rosenthal, C., Costa, G., Spangler, J., Sims, E. H., et al. (2013). Rapid 16S rRNA Next-Generation Sequencing of Polymicrobial Clinical Samples for Diagnosis of Complex Bacterial Infections. PLoS ONE, 8(5), e65226. http://doi.org/10.1371/journal.pone.0065226

Sheward, D. J., Murrell, B., & Williamson, C. (2012). Degenerate Primer IDs and the birthday problem. Proceedings of the National Academy of Sciences, 109(21), E1330-E1330.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). Pear: a fast and accurate Illumina paired-end read merger. Bioinformatics, 30(5):614.